

A Note on Amemiya's Form of the
Weighted Least Squares Estimator

Roger Koenker
Christopher L. Skeels
A. H. Welsh

The Library of the
DEC 5 1990

University of Illinois
at Urbana-Champaign



College of Commerce and Business Administration
Bureau of Economic and Business Research
University of Illinois Urbana-Champaign

BEBR

FACULTY WORKING PAPER NO. 90-1691

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

September 1990

A Note on Amemiya's Form of the Weighted
Least Squares Estimator

Roger Koenker
Department of Economics
University of Illinois

Christopher L. Skeels
A. H. Welsh
Department of Statistics
Faculty of Economics and Commerce
Australian National University

Digitized by the Internet Archive
in 2011 with funding from
University of Illinois Urbana-Champaign

ABSTRACT

Amemiya's estimator is a weighted least squares estimator of the regression coefficients in a linear model with heteroskedastic errors. It is attractive because the heteroscedasticity is not parameterised and the weights (which depend on the error covariance matrix) are estimated nonparametrically. In this paper, we obtain an asymptotic expansion for Amemiya's form of the weighted least squares estimator. We use this expansion to discuss the effect of estimating the weights, the effect of the number of iterations and the effect of the choice of the initial estimate. We also discuss the special case of normally distributed errors and clarify the special consequences of assuming normality.

1. Introduction

Econometric modelling is frequently complicated by heterogeneous variability in the stochastic component of the model. Such heteroscedasticity, arises in almost all fields; for examples see Carroll and Ruppert (1988). It is always possible, of course, to ignore the heteroscedasticity and proceed with a standard analysis, but substantial gains in efficiency are possible if we incorporate information about the heteroscedasticity into the analysis. One approach is to model the heteroscedasticity by introducing an explicit parametric model for the scale of the stochastic component of the model. This approach has been explored in considerable detail; again see Carroll and Ruppert (1988) for a recent survey. It can, however, be prohibitively difficult to parametrize heteroscedasticity. In practice purely empirical models are difficult to identify, and there may be no theoretical motivation for a particular structural model. Economic theory is rich in models for conditional means but meagre as a source of models for scale. In this paper, therefore, we will consider an approach suggested by Amemiya (1983) which attempts to deal with heteroscedasticity without introducing an explicit parametric model. This approach is closely allied with the work of Eicker (1963), and White (1982) on consistent covariance matrix estimation and Chamberlain (1982) on method of moments estimation.

Consider the heteroscedastic linear model

$$y = X\beta + u, \quad (1.1)$$

where X is an $n \times p$ matrix of known constants with rows denoted x_j^T , $j = 1, \dots, n$, β is a p -vector of unknown parameters and $u = (u_1, \dots, u_n)^T$ is an n -vector of independent random variables with $Eu_j = 0$, $Eu_j^2 = \sigma_j^2$, $Eu_j^3 = \mu_{3j}$ and $Eu_j^4 = \mu_{4j} < \infty$. The regression parameter β is the parameter of interest while $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ is regarded as an arbitrary n dimensional nuisance parameter. In the classical linear model we take u_1, \dots, u_n to be identically distributed so that $\Sigma = \sigma^2 I$.

The weighted least squares estimator is widely used for estimating the regression parameter in heteroscedastic linear models. Notice that when Σ is known, premultiplying (1.1) by $\Sigma^{-1/2}$ yields a classical linear model for which the least squares estimator is

$$\hat{\beta}_\Sigma = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} y.$$

When Σ is unknown $\hat{\beta}_{\Sigma}$ cannot be computed but we may be able to substitute an appropriate $\hat{\Sigma}$ for Σ to obtain

$$\hat{\beta}_{\hat{\Sigma}} = (X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1} y.$$

Since Σ is not parametrized, an appropriate $\hat{\Sigma}$ is obtained by setting $\hat{\Sigma} = \text{diag}(r_1^2, \dots, r_n^2)$, where $r = Y - X\hat{\beta}_{(0)}$ is the vector of residuals from an initial estimator $\hat{\beta}_{(0)}$ of β . Notice that $\hat{\Sigma}$ is in fact an estimator of $\text{diag}(u_1^2, \dots, u_n^2)$ rather than of Σ . Although we actually need to estimate $n^{-1}X^T\Sigma^{-1}X$ and $n^{-1}X^T\Sigma^{-1}y$ rather than Σ , it turns out that we cannot estimate $n^{-1}X^T\Sigma^{-1}X$ unless we can estimate Σ . As we have only n observations with which to estimate the n parameters in Σ we cannot construct a consistent estimator of Σ . However, there is a convenient reformulation of $\hat{\beta}_{\Sigma}$ which enables us to overcome this difficulty. Let V be an $n \times (n-p)$ matrix of constants such that (X, V) is a nonsingular $n \times n$ matrix and $V^T X = 0$. i.e. the columns of V are orthogonal to those of X . If we let $\mathcal{R}(X)$ denote the subspace of \mathbb{R}^n spanned by the rows of X and $\mathcal{R}(X)^{\perp}$ denote its orthogonal complement, we have trivially that $\mathcal{R}(\Sigma^{-1/2}X)^{\perp} = \mathcal{R}(X)^{\perp} = \mathcal{R}(V) = \mathcal{R}(\Sigma^{1/2}V)$. Now $I - \Sigma^{-1/2}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1/2}$ projects \mathbb{R}^n onto $\mathcal{R}(\Sigma^{-1/2}X)^{\perp}$ and $\Sigma^{1/2}V(V^T\Sigma V)^{-1}V^T\Sigma^{1/2}$ projects \mathbb{R}^n onto $\mathcal{R}(\Sigma^{1/2}V)$ and this projection is unique (e.g. Seber, 1977, p394), so we have $I - \Sigma^{-1/2}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1/2} = \Sigma^{1/2}V(V^T\Sigma V)^{-1}V^T\Sigma^{1/2}$. Thus

$$\begin{aligned} \hat{\beta}_{\Sigma} &= (X^T X)^{-1} X^T y - (X^T X)^{-1} X^T \Sigma^{1/2} \{I - \Sigma^{-1/2} X (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1/2}\} \Sigma^{-1/2} y \\ &= (X^T X)^{-1} X^T y - (X^T X)^{-1} X^T \Sigma^{1/2} \{\Sigma^{1/2} V (V^T \Sigma V)^{-1} V^T \Sigma^{1/2}\} \Sigma^{-1/2} y \\ &= (X^T X)^{-1} \{X^T - X^T \Sigma V (V^T \Sigma V)^{-1} V^T\} y \\ &= \hat{\beta}_I - (X^T X)^{-1} X^T \Sigma V (V^T \Sigma V)^{-1} V^T y \end{aligned}$$

which involves Σ rather than Σ^{-1} . At least so far as analysis is concerned, there is a slight further difficulty caused by the fact that the dimensions of $n^{-1}V^T\Sigma V$ and $n^{-1}X^T\Sigma V$ increase with n . Amemiya (1983) therefore suggested that we replace V by an $n \times q$ matrix W with a fixed $q < n-p$ of the columns of V . Replacing Σ by $\hat{\Sigma}$, we obtain Amemiya's estimator

$$\hat{\beta}_{(1)} = \hat{\beta}_I - (X^T X)^{-1} X^T \hat{\Sigma} W (W^T \hat{\Sigma} W)^{-1} W^T y. \quad (1.2)$$

If $\Sigma = \sigma^2 I$ is known, $\hat{\beta}_{(1)} = \hat{\beta}_I$. It is obvious that in replacing V by W we are neglecting some of the structure of Σ . Nonetheless, Amemiya (1983) showed that this estimator is always more efficient than the least squares estimator $\hat{\beta}_I$ and Balestra (1983) showed that it can be as efficient as $\hat{\beta}_{\Sigma}$; in particular, if there are only q different variances in Σ , a judicious choice of W makes Amemiya's estimator equal to $\hat{\beta}_{\Sigma}$. The general issue of how to choose W has not been addressed. Nor has the possibility of allowing q to diverge to infinity at a slower rate than n . These interesting issues are beyond the scope of the present paper and will not be pursued here. Our purpose is rather to obtain an expansion for $\hat{\beta}_{(1)}$ which enables us to examine the effect of using $\hat{\Sigma}$ rather than Σ in the estimator, the effect of the number of iterations and the effect of the choice of the initial estimate. This work complements that of Carroll, Wu and Ruppert (1988) and Rothenberg (1984) on the effect on weighted least squares of fitting parametric models for Σ and extends that of Fuller and Rao (1978) on the replicated case by relaxing the assumption of normal errors.

2. Theoretical Results

Our main result is a higher order asymptotic expansion for $\hat{\beta}_{(1)}$ including terms of order $n^{-3/2}$ in probability. The expansion requires conditions on various sums and matrices involving X , W and the moments of the u_j 's which are stated in the Section 3. We also require a condition on the initial estimator $\hat{\beta}_{(0)}$. In particular, we suppose that $\hat{\beta}_{(0)}$ satisfies

$$\hat{\beta}_{(0)} - \beta = n^{-1} C^{-1} D^T \Psi(u) + o_p(n^{-1/2}), \quad (2.1)$$

for some $p \times p$ nonsingular matrix $C = O(1)$, some $n \times p$ matrix D and some vector function $\Psi(u) = (\psi(u_1), \dots, \psi(u_n))^T$, where $n^{-1} C^{-1} D^T \Psi(u) = O_p(n^{-1/2})$. It is convenient to set

$$A = n^{-1} W^T \Sigma W \quad \text{and} \quad M = X - n^{-1} W A^{-1} W^T \Sigma X.$$

Then we show in Section 3 that

$$\begin{aligned} \hat{\beta}_{(1)} - \beta &= n^{-1/2} Z_{1n} + n^{-1} Z_{2n} + n^{-3/2} Z_{3n} (\hat{\beta}_{(0)} - \beta) + o_p(n^{-3/2}), \\ &= n^{-1/2} Z_{1n} + n^{-1} Z_{2n} + n^{-3/2} Z_{3n} (n^{-1} C^{-1} D^T \Psi(u)) + o_p(n^{-3/2}), \end{aligned} \quad (2.2)$$

where $n^{-1/2}Z_{1n} = (X^T X)^{-1} M^T u$ and $Z_{tn} = O_p(1)$, $t = 1, 2, 3$. Here $Z_{3n}(\cdot): \mathbb{R}^P \rightarrow \mathbb{R}^P$ is a function of the initial estimator whereas Z_{1n} and Z_{2n} are not. If Σ were known, we would have the identity

$$\hat{\beta}_{(1)} - \beta = n^{-1/2} Z_{1n} = (X^T X)^{-1} M^T u,$$

and

$$\begin{aligned} \text{Var } \hat{\beta}_{(1)} &= n^{-1} E Z_{1n} Z_{1n}^T \\ &= (X^T X)^{-1} M^T \Sigma M (X^T X)^{-1} \\ &= \text{Var } \hat{\beta}_I - (X^T X)^{-1} X^T \Sigma W A^{-1} W \Sigma X (X^T X)^{-1}. \end{aligned}$$

which, incidentally, proves that $\text{Var } n^{-1/2} Z_{1n} \leq \text{Var } \hat{\beta}_I$. When Σ is unknown, we have (2.2)

and, proceeding formally, the moment expansions

$$E \hat{\beta}_{(1)} - \beta = n^{-1} E Z_{2n} + o(n^{-1})$$

and

$$\text{Var } \hat{\beta}_{(1)} = n^{-1} E Z_{1n} Z_{1n}^T + n^{-2} T(\hat{\beta}_{(0)} - \beta) + o(n^{-2}), \quad (2.3)$$

where

$$\begin{aligned} T(\hat{\beta}_{(0)} - \beta) &= E Z_{2n} Z_{1n}^T + E Z_{1n} Z_{2n}^T + E Z_{2n} Z_{2n}^T - E Z_{2n} E Z_{2n}^T \\ &\quad + E Z_{3n}(\hat{\beta}_{(0)} - \beta) Z_{1n}^T + E Z_{1n} Z_{3n}(\hat{\beta}_{(0)} - \beta)^T. \end{aligned}$$

It is instructive to write $T = T_1 + T_2(\hat{\beta}_{(0)} - \beta)$, where $T_2(\cdot)$ is a function of $\hat{\beta}_{(0)} - \beta$ and T_1 is not. It follows from the results in Section 3 that

$$\begin{aligned} T_1 &= -n^{-1} (X^T X)^{-1} \sum_{j=1}^n m_j m_j^T w_j^T A^{-1} w_j (\mu_{4j} - \sigma_j^4) (X^T X)^{-1} \\ &\quad + 3 n^{-2} (X^T X)^{-1} \sum_{j=1}^n \sum_{k=1}^n m_j m_k^T (w_j^T A^{-1} w_k)^2 \mu_{3j} \mu_{3k} (X^T X)^{-1} \\ &\quad + 2 n^{-2} (X^T X)^{-1} \sum_{j=1}^n \sum_{k=1}^n m_j m_j^T w_j^T A^{-1} w_j w_k^T A^{-1} w_k \mu_{3j} \mu_{3k} (X^T X)^{-1} \end{aligned} \quad (2.4)$$

and

$$T_2(n^{-1}C^{-1}D^T\Psi(u)) = 4n^{-2}(X^TX)^{-1} \sum_{j=1}^n m_j m_j^T w_j^T A^{-1} W^T G D C^{-1} x_j \sigma_j^2 (X^TX)^{-1} \quad (2.5)$$

$$+ 4n^{-2}(X^TX)^{-1} \sum_{j=1}^n m_j x_j^T \{w_j^T A^{-1} w_j \sigma_j^2 - n^{-1} x_j^T C^{-1} D^T G W A^{-1} w_j\} C^{-1} D^T G M (X^TX)^{-1},$$

where $G = \text{diag}(Eu_1\Psi(u_1), \dots, Eu_n\Psi(u_n))$.

It is perhaps worth noting that, with considerable work, higher order terms in the above expansions could be obtained. However, the above expansions contain sufficient terms to capture the dominant effect of the initial estimator. Moreover, Carroll, Wu and Ruppert (1988) found that the conclusions drawn from examining expansions of this order seem to reflect, at least qualitatively, the findings from simulation studies.

The contribution of the initial estimator to Amemiya's estimator (1.2) is of order $n^{-3/2}$ in probability and affects the second term in the expansion of the asymptotic variance of Amemiya's estimator. We can iterate the procedure by using $\hat{\beta}_{(1)}$ as a new initial estimator, calculating $\hat{\beta}_{(2)}$ etc. Identifying $n^{-1/2}Z_{1n} = (X^TX)^{-1}M^Tu$ with $n^{-1}C^{-1}D^T\Psi(u)$, $nC = X^TX$, $D = M$ and $\Psi(u) = u$, we find that for $c \geq 2$, (2.2) becomes

$$\hat{\beta}_{(c)} - \beta = n^{-1/2}Z_{1n} + n^{-1}Z_{2n} + n^{-3/2}Z_{3n}(n^{-1/2}Z_{1n}) + o_p(n^{-3/2}),$$

and (2.3) becomes

$$\text{Var } \hat{\beta}_{(c)} = n^{-1}EZ_{1n}Z_{1n}^T + n^{-2}\{T_1 + T_2(n^{-1/2}Z_{1n})\} + o(n^{-2}),$$

Thus iteration reduces the contribution of the initial estimator to a smaller order than $n^{-3/2}$ in probability and the first two terms of the asymptotic variance stabilise after two iterations. Carroll, Wu and Ruppert (1988) obtained a similar result when the parametric model for Σ does not depend on $X\beta$ but that an extra iteration is required to achieve stability when the model for Σ depends on $X\beta$.

It is not always straightforward to draw general conclusions from the expansions (2.4) and (2.5) so it is worth considering the simple special case that the u_i 's are identically distributed with a symmetric distribution so that $\Sigma = \sigma^2 I$, $\mu_{3j} = 0$ and $\mu_{4j} = \mu_4$. Notice that here we are examining the consequences of proceeding as though we had a heteroscedastic model when in fact we do not. In this case $M = X$ and (2.4) and (2.5) become

$$T_1 = -(\mu_4 - \sigma^4) n^{-1}(X^T X)^{-1} \sum_{j=1}^n x_j x_j^T w_j^T A^{-1} w_j (X^T X)^{-1} \quad (2.6)$$

and

$$\begin{aligned} T_2(n^{-1} C^{-1} D^T \Psi(u)) &= 4\sigma^2 E u_1 \psi(u_1) n^{-2} (X^T X)^{-1} \sum_{j=1}^n x_j x_j^T \{w_j^T A^{-1} W^T D C^{-1} x_j \\ &\quad + C^{-1} D^T X w_j^T A^{-1} w_j\} (X^T X)^{-1} \\ &\quad - 4 \{E u_1 \psi(u_1)\}^2 n^{-3} (X^T X)^{-1} \sum_{j=1}^n x_j x_j^T C^{-1} D^T X x_j^T C^{-1} D^T W A^{-1} w_j (X^T X)^{-1}. \end{aligned} \quad (2.7)$$

Interestingly, Carroll, Wu and Ruppert (1988) found that using the least squares estimator $\hat{\beta}_I$ as the initial estimator reduced the number of iterations for the covariance to stabilise by one in each case. This is not in general true for Amemiya's estimator. However, when the u_i 's are identically and symmetrically distributed the least squares estimator $\hat{\beta}_I$ satisfies (2.1) with $nC = X^T X$, $D = X$ and $\psi(u) = u$ and, for $c \geq 2$, $\hat{\beta}_{(c-1)}$ satisfies (2.1) with $nC = X^T X$, $D = M = X$ and $\psi(u) = u$ so that

$$T_2(\hat{\beta}_I - \beta) = T_2(\hat{\beta}_{(c-1)} - \beta) = 4\sigma^4 n^{-1} (X^T X)^{-1} \sum_{j=1}^n x_j x_j^T w_j^T A^{-1} w_j (X^T X)^{-1}, \quad c \geq 2. \quad (2.8)$$

Thus in this particular case, starting with the least squares estimator results in a stable covariance after only one iteration.

Carroll, Wu and Ruppert (1988) show further that there may be advantages to using a robust initial estimator. Suppose we use the M-estimator β^* obtained by solving $\sum_{j=1}^n x_j \psi((y_j - x_j^T \beta)/\hat{\omega}) = 0$, where $\hat{\omega}$ is a consistent estimate of some scale functional ω which need not equal σ . If the u_i 's are identically distributed with a symmetric distribution, β^* satisfies (2.1) with $nC = \omega^{-1} E \psi'(u_1/\omega) X^T X$ and $D = X$ so that

$$T_2(\beta^* - \beta) = 4\sigma^2 \frac{E u_1 \psi(u_1/\omega)}{\omega^{-1} E \psi'(u_1/\omega)} n^{-1} (X^T X)^{-1} \sum_{j=1}^n x_j x_j^T w_j^T A^{-1} w_j (X^T X)^{-1}. \quad (2.9)$$

Since $\sum_{j=1}^n x_j x_j^T w_j^T A^{-1} w_j$ is nonnegative definite, a comparison of (2.8) and (2.9) shows that

$\hat{\beta}_{(1)}$ based on an M-estimator has a smaller covariance (up to terms of order n^{-2}) than $\hat{\beta}_{(1)}$

based on the least squares estimator or, indeed, on the iterated stable estimator $\hat{\beta}_{(c)}$, $c \geq 2$, whenever

$$Eu_1\psi(u_1/\omega) < \sigma^2\omega^{-1}E\psi'(u_1/\omega). \quad (2.10)$$

Note that more generally when the u_i 's have arbitrary symmetric distributions β^* satisfies (2.1) with $nC = \omega^{-1}X^T \text{diag}(E\psi'(u_1/\omega), \dots, E\psi'(u_n/\omega))X$ and $D = X$ so that we can write down expansions for this case too. Moreover, we can also drop the symmetry assumption but at the cost of a slightly more sophisticated analysis.

We can also examine the effect of including $\hat{\Sigma}$ in our analysis when it is not actually required in the identically distributed symmetric case. Since

$$\begin{aligned} T(\beta^* - \beta) &= T_1 + T_2(\beta^* - \beta) \\ &= - \left\{ \mu_4 - \sigma^4 - 4\sigma^2 \frac{Eu_1\psi(u_1/\omega)}{\omega^{-1}E\psi'(u_1/\omega)} \right\} n^{-1}(X^TX)^{-1} \sum_{j=1}^n x_j x_j^T w_j^T A^{-1} w_j (X^TX)^{-1} \end{aligned}$$

and

$$T(\hat{\beta}_1 - \beta) = T(\hat{\beta}_{(c-1)} - \beta) = -\{\mu_4 - 5\sigma^4\} n^{-1}(X^TX)^{-1} \sum_{j=1}^n x_j x_j^T w_j^T A^{-1} w_j (X^TX)^{-1}, \quad c \geq 2,$$

we see that for near normal distributions with $\kappa = \mu_4/\sigma^4 < 5$, including $\hat{\Sigma}$ when it is not actually required causes an increase in the covariance compared to when $\Sigma = \sigma^2 I$ is known. However, for long-tailed distributions with $\kappa > 5$, including $\hat{\Sigma}$ actually reduces the covariance (up to $O(n^{-2})$) compared to when $\Sigma = \sigma^2 I$ is known. The same result was found in Carroll, Wu and Ruppert (1988). One possible explanation is that when we have long-tailed distributions we obtain some large residuals and weighted least squares estimators downweight the observations corresponding to these residuals so that we actually get a kind of robustness effect.

Finally, consider the particular case where u has a multivariate normal distribution. Rothenberg (1984) has examined the special case where Σ depends on a finite dimensional parameter θ which is not a function of β . He assumes that $\hat{\Sigma}$ is formed from estimates $\hat{\theta}$ which are even functions of u and also do not depend on β . Given the closure of the multivariate normal distribution under linear transformations, this last condition implies that $\hat{\theta}$ is an even function of the projection of y onto the orthogonal complement of the column space of X . That is, the initial estimator $\hat{\beta}_{(0)}$ will be of the form $\hat{\beta}_{(0)} = (\bar{X}^T Q \bar{X})^{-1} \bar{X}^T Q y$, where Q is

an arbitrary positive definite matrix not depending on β and \bar{X} is any matrix which spans the column space of X . Then $\hat{\theta}$ is obtained as an even function of the resulting residuals. He found that including $\hat{\Sigma}$ increases the covariance compared to when Σ is known, that the number of iterations and the choice of Q do not matter. Note that for normal u_1 an integration by parts implies that $E u_1 \psi(u_1/\omega) = \sigma^2 \omega^{-1} E \psi'(u_1/\omega)$ and so (2.10) cannot hold. But in non-normal models, however, choosing ψ so that (2.10) holds, we can actually decrease the covariance (up to n^{-2}) compared to when Σ is known. Moreover, even if we restrict attention to linear initial estimators we find that the number of iterations does matter. Here on setting $\bar{X} = X$, the most plausible choice for X , $\hat{\beta}_{(0)} = (X^T Q X)^{-1} X^T Q y$ satisfies (2.1) with $nC = X^T Q X$, $D = QX$ and $\psi(u) = u$ so $G = \Sigma$ and (2.5) becomes

$$\begin{aligned} T_2((X^T Q X)^{-1} X^T Q u) &= 4n^{-1} (X^T X)^{-1} \sum_{j=1}^n m_j m_j^T w_j^T A^{-1} W^T \Sigma Q X (X^T Q X)^{-1} x_j \sigma_j^2 (X^T X)^{-1} \\ &\quad + 4n^{-1} (X^T X)^{-1} \sum_{j=1}^n m_j x_j^T w_j^T A^{-1} w_j \sigma_j^2 (X^T Q X)^{-1} X^T Q \Sigma M (X^T X)^{-1} \\ &\quad - 4n^{-1} (X^T X)^{-1} \sum_{j=1}^n m_j x_j^T (X^T Q X)^{-1} X^T Q \Sigma M x_j^T (X^T Q X)^{-1} X^T Q \Sigma W A^{-1} w_j (X^T X)^{-1} \end{aligned}$$

which depends on Q . However, in the identically distributed case ($\Sigma = \sigma^2 I$), the number of iterations and the choice of Q do not matter as $M = X$ and (2.7) becomes

$$T_2((X^T Q X)^{-1} X^T Q u) = 4\sigma^4 n^{-1} (X^T X)^{-1} \sum_{j=1}^n x_j x_j^T w_j^T A^{-1} w_j (X^T X)^{-1}$$

which does not depend on Q .

3. Proofs

In this section we give a formal proof of the expansion (2.2), obtain expressions for Z_{in} , $t = 1, 2, 3$, and then calculate formally the moments which appear in T_1 and $T_2(\cdot)$.

To prove (2.2) suppose that (2.1) holds and that, with $M = X - W(W^T \Sigma W)^{-1} W^T \Sigma X$,

i) $n^{-1} X^T X$ and $n^{-1} W^T \Sigma W$ converge to nonsingular limits,

and

ii) $n^{-1} X^T \Sigma X = O(1)$,

$n^{-1} W^T \Sigma X = O(1)$,

$n^{-1} \sum_{j=1}^n |w_j w_j^T| |x_j|^2 = O(1)$,

$n^{-1} \sum_{j=1}^n |m_j w_j^T| |x_j|^2 = O(1)$

$$n^{-1} \sum_{j=1}^n (w_j w_j^T) * (w_j w_j^T) \mu_{4j} = O(1) \quad n^{-1} \sum_{j=1}^n (m_j w_j^T) * (m_j w_j^T) \mu_{4j} = O(1)$$

$$n^{-1} \sum_{j=1}^n w_j w_j^T w_{jk}^2 w_{jl}^2 \sigma_j^2 = O(1), \quad 1 \leq k, l \leq q,$$

$$n^{-1} \sum_{j=1}^n x_j x_j^T w_{jk}^2 m_{jl}^2 \sigma_j^2 = O(1), \quad 1 \leq k \leq q, 1 \leq l \leq p,$$

hold. (Here $*$ denotes the Hadamard product of two matrices.)

First note that as $W^T X = 0$, we can write

$$\begin{aligned} \hat{\beta}_{(1)} &= \hat{\beta}_I - (X^T X)^{-1} X^T \hat{\Sigma} W (W^T \hat{\Sigma} W)^{-1} W^T Y \\ &= \beta + (X^T X)^{-1} \{X^T - X^T \hat{\Sigma} W (W^T \hat{\Sigma} W)^{-1} W^T\} u. \end{aligned} \quad (3.1)$$

To preserve notation let

$$G_1 = \text{diag}(u_1^2 - \sigma_1^2, \dots, u_n^2 - \sigma_n^2)$$

$$G_2 = \text{diag}(u_1 x_1^T (\hat{\beta} - \beta), \dots, u_n x_n^T (\hat{\beta} - \beta))$$

and

$$G_3 = \text{diag}(\{x_1^T (\hat{\beta} - \beta)\}^2, \dots, \{x_n^T (\hat{\beta} - \beta)\}^2).$$

Then, squaring the residuals, we obtain

$$\begin{aligned} n^{-1} X^T \hat{\Sigma} W &= n^{-1} X^T \Sigma W + n^{-1} X^T \text{diag}(r_1^2 - \sigma_1^2, \dots, r_n^2 - \sigma_n^2) W \\ &= n^{-1} X^T \Sigma W + n^{-1} X^T G_1 W - 2n^{-1} X^T G_2 W + n^{-1} X^T G_3 W \end{aligned} \quad (3.2)$$

and, similarly,

$$n^{-1} W^T \hat{\Sigma} W = n^{-1} W^T \Sigma W + n^{-1} W^T G_1 W - 2n^{-1} W^T G_2 W + n^{-1} W^T G_3 W.$$

Notice that when $\hat{A} - A = O_p(n^{-1/2})$ we have

$$\hat{A}^{-1} = A^{-1} - A^{-1}(\hat{A} - A)A^{-1} + A^{-1}(\hat{A} - A)A^{-1}(\hat{A} - A)A^{-1} + O_p(n^{-3/2})$$

so that with $\hat{A} = n^{-1} W^T \hat{\Sigma} W$ and $A = n^{-1} W^T \Sigma W$, we obtain

$$\begin{aligned} n(W^T \hat{\Sigma} W)^{-1} &= A^{-1} - A^{-1} n^{-1} W^T G_1 W A^{-1} + A^{-1} n^{-1} W^T G_1 W A^{-1} W^T G_1 W A^{-1} \\ &\quad + 2A^{-1} W^T G_2 W A^{-1} - A^{-1} n^{-1} W^T G_3 W A^{-1} + O_p(n^{-3/2}). \end{aligned} \quad (3.3)$$

Substituting (3.2), (3.3) and (2.1) into (3.1) yields

$$\hat{\beta}_{(1)} - \beta = Z_{1n} + Z_{2n} + Z_{3n}(\hat{\beta} - \beta) + o_p(n^{-3/2}),$$

where,

$$Z_{1n} = (X^T X)^{-1} M^T u = O_p(n^{-1/2})$$

$$Z_{2n} = -n^{-1} (X^T X)^{-1} M^T G_1 W A^{-1} W^T u = O_p(n^{-1})$$

and

$$\begin{aligned} Z_{3n}(\hat{\beta}_{(0)} - \beta) &= n^{-2} (X^T X)^{-1} M^T G_1 W A^{-1} W^T G_1 W A^{-1} W^T u \\ &\quad + 2n^{-1} (X^T X)^{-1} M^T G_2 W A^{-1} W^T u - n^{-1} (X^T X)^{-1} M^T G_3 W A^{-1} W^T u \\ &= n^{-2} (X^T X)^{-1} M^T G_1 W A^{-1} W^T G_1 W A^{-1} W^T u \\ &\quad + 2n^{-2} (X^T X)^{-1} M^T \text{diag}(u_1 x_1^T C^{-1} D \Psi(u), \dots, u_n x_n^T C^{-1} D \Psi(u)) W A^{-1} W^T u \\ &\quad - n^{-2} (X^T X)^{-1} M^T \text{diag}(x_1^T C^{-1} D \Psi(u), \dots, x_n^T C^{-1} D \Psi(u))^2 W A^{-1} W^T u \\ &= O_p(n^{-3/2}). \end{aligned}$$

Now writing $D^T = (d_1, \dots, d_n)$, $G = \text{diag}(Eu_1 \psi(u_1), \dots, Eu_n \psi(u_n))$ and proceeding formally,

$$EZ_{1n} = 0;$$

$$EZ_{1n} Z_{1n}^T = (X^T X)^{-1} M \Sigma M^T (X^T X)^{-1};$$

$$EZ_{2n} = -n^{-1} (X^T X)^{-1} \sum_{j=1}^n m_j w_j^T A^{-1} w_j \mu_{3j};$$

$$EZ_{2n} Z_{1n}^T = -n^{-1} (X^T X)^{-1} \sum_{j=1}^n m_j m_j^T w_j^T A^{-1} w_j (\mu_{4j} - \sigma_j^4) (X^T X)^{-1}$$

$$EZ_{2n} Z_{2n}^T = n^{-2} (X^T X)^{-1} \sum_{j=1}^n \sum_{k=1}^n m_j m_k^T w_j^T A^{-1} w_j w_k^T A^{-1} w_k \mu_{3j} \mu_{3k} (X^T X)^{-1}$$

$$+ n^{-2} (X^T X)^{-1} \sum_{j=1}^n \sum_{k=1}^n m_j m_k^T (w_j^T A^{-1} w_k)^2 \mu_{3j} \mu_{3k} (X^T X)^{-1}$$

$$+ n^{-1} (X^T X)^{-1} \sum_{j=1}^n m_j m_j^T w_j^T A^{-1} w_j (\mu_{4j} - \sigma_j^4) (X^T X)^{-1} + O(n^{-3})$$

$$\begin{aligned}
EZ_{3n}Z_{1n}^T &= n^{-2}(X^TX)^{-1} \sum_{j=1}^n \sum_{k=1}^n m_j m_k^T w_j^T A^{-1} w_k w_k^T A^{-1} w_j \mu_{3j} \mu_{3k} (X^TX)^{-1} \\
&+ n^{-2}(X^TX)^{-1} \sum_{j=1}^n \sum_{k=1}^n m_j m_k^T (w_j^T A^{-1} w_k)^2 \mu_{3j} \mu_{3k} (X^TX)^{-1} \\
&+ 2n^{-2}(X^TX)^{-1} \sum_{j=1}^n m_j m_j^T w_j^T A^{-1} W^T GDC^{-1} x_j \sigma_j^2 (X^TX)^{-1} \\
&+ 2n^{-2}(X^TX)^{-1} \sum_{j=1}^n m_j x_j^T \{ w_j^T A^{-1} w_j \sigma_j^2 - n^{-1} x_j^T C^{-1} D^T GWA^{-1} w_j \} C^{-1} D^T GM (X^TX)^{-1} \\
&+ O(n^{-3}),
\end{aligned}$$

$$\text{as } A^{-1}W^T \Sigma M = A^{-1}W^T \Sigma X - n^{-1}A^{-1}W^T \Sigma W A^{-1}W^T \Sigma X = A^{-1}W^T \Sigma X - A^{-1}W^T \Sigma X = 0.$$

Finally,

$$\begin{aligned}
EZ_{2n}Z_{1n}^T + EZ_{1n}Z_{2n}^T + EZ_{2n}Z_{2n}^T - EZ_{2n}EZ_{2n}^T + EZ_{3n}(\hat{\beta} - \beta)Z_{1n}^T + EZ_{1n}Z_{3n}^T(\hat{\beta} - \beta) \\
= -n^{-1}(X^TX)^{-1} \sum_{j=1}^n m_j m_j^T w_j^T A^{-1} w_j (\mu_{4j} - \sigma_j^4)(X^TX)^{-1} \\
+ 3n^{-2}(X^TX)^{-1} \sum_{j=1}^n \sum_{k=1}^n m_j m_k^T (w_j^T A^{-1} w_k)^2 \mu_{3j} \mu_{3k} (X^TX)^{-1} \\
+ 2n^{-2}(X^TX)^{-1} \sum_{j=1}^n \sum_{k=1}^n m_j m_j^T w_j^T A^{-1} w_j w_k^T A^{-1} w_k \mu_{3j} \mu_{3k} (X^TX)^{-1} \\
+ 4n^{-2}(X^TX)^{-1} \sum_{j=1}^n m_j m_j^T w_j^T A^{-1} W^T GDC^{-1} x_j \sigma_j^2 (X^TX)^{-1} \\
+ 4n^{-2}(X^TX)^{-1} \sum_{j=1}^n m_j x_j^T \{ w_j^T A^{-1} w_j \sigma_j^2 - n^{-1} x_j^T C^{-1} D^T GWA^{-1} w_j \} C^{-1} D^T GM (X^TX)^{-1}.
\end{aligned}$$

4. Numerical Results

We performed a limited simulation experiment to examine some of the predictions of the asymptotics, the results of which are presented in Table 1. Using a sample size of 50 we fitted a linear regression through the origin with $X \sim N(0, 25)$ and the coefficient on x set to unity. Although not reported, other sample sizes were examined with improved performance, measured in terms of mean squared error, as the sample size increased and inferior performance for smaller sample sizes. The disturbance term had zero mean but its distribution differed from case to case. The M-estimator chosen as an initial estimator was that proposed by Huber (1964) with c , using the notation of Amemiya (1985, equation 2.3.2), chosen to be 1.345.

Some experimentation suggested that the results obtained were relatively robust to the choice of c . In constructing the weighted least squares estimator, W was chosen (initially) to be the first column of $P_X = I_n - X(X^T X)^{-1} X^T$. In what follows we shall denote the iterated weighted least squares estimator, using ordinary least squares as an initial estimator, by β_{ls} ; β_m shall denote its analogue based on the M-estimator. This notation suppresses the number of iterations used in the estimation process. In Table 1, mean squared errors are reported for estimators involving one through five iterations, inclusive. All results are based on 1000 replications.

As a bench-mark we can compare the performance of β_{ls} and β_m when the disturbances of the model are $u_i \sim N(0, 1)$, $i = 1, \dots, n$ (experiment 1), and when $u_i \sim t(5)$, $i = 1, \dots, n$ (experiment 2). In both experiments the disturbances are homoscedastic. In the latter experiment $\kappa = 9$ and, as predicted, β_m performs better than β_{ls} although, as in experiment 1, there is little to choose between them. One common feature of the two sets of results is that nothing appears to be gained by iterating. Indeed mean squared error seems to increase with the number of iterations. There was some evidence to suggest that the mean squared error converged to some finite value, usually within four to seven iterations.

=====

Table 1 about here

=====

Experiments 3 and 4 repeat the first two experiments but with $u_i \sim N(0, i)$, $i = 1, \dots, n$ and $u_i \sim i^{1/2} v_i$, $v_i \sim t(5)$, $i = 1, \dots, n$, respectively. That is, these experiments consider heteroscedastic models with the scale of the disturbance increasing with the index. The most noticeable feature of these results is the dramatic decline in the performance of the estimators relative to that for the homoscedastic models. In experiment 3 we see that, for $\kappa < 5$, there remains little to choose between the two estimators. In contrast, the results of experiment 4 suggest that as the error distribution becomes increasingly leptokurtotic there are benefits in using a robust initial estimator.

As indicated in the introduction, no effort has been devoted finding the optimal W for the estimator although Balestra (1983) has shown that in certain special situations there may exist such a choice. Nevertheless some investigation of the effect of different choices for W was made by using different columns of P_X in the construction of the estimators. The worst

case that was found is presented as experiment 5. It is evident from the results the performance of both estimators is dramatically worse than for the other experiments. Further, the mean squared errors are oscillating quite violently. While not entirely understood, it may be that these results are driven by the inversion of an ill-conditioned matrix, there is enough evidence to suggest that these weighted least squares estimators are sensitive to the choice of W . This remains a topic for further research.

* The authors would like to thank Trevor Breusch, Jose Machado and Terry O'Neill for helpful discussions. The usual caveat applies.

References

- Amemiya, T., 1983, Partially generalised least squares and two-stage least squares estimators, *Journal of Econometrics* 23, 275-283.
- Amemiya, T., 1985, *Advanced Econometrics* (Basil Blackwell).
- Balestra, P., 1983, A note on Amemiya's partially generalised least squares, *Journal of Econometrics* 23, 285-290.
- Carroll, R.J. and Ruppert, D., 1988, *Transformation and weighting in regression* (Chapman and Hall, New York).
- Carroll, R.J., Wu, C.F.J. and Ruppert, D., 1988, The effect of estimating weights in weighted least squares, *Journal of the American Statistical Association* 83, 1045-1054.
- Chamberlain, G., 1982, Multivariate regression models for panel data, *Journal of Econometrics* 18, 5-46.
- Eicker, F., 1963, Asymptotic normality and consistency of the least squares estimators for families of linear regressions, *Annals of Mathematical Statistics* 34, 447-456.
- Fuller, W.A. and Rao, J.N.K., 1978, Estimation for a linear regression model with unknown diagonal covariance matrix, *Annals of Statistics* 6, 1149-1158.
- Huber, P.J., 1964, Robust estimation of a location parameter, *Annals of Mathematical Statistics* 35, 73-101.
- Rothenberg, T.J., 1984, Approximate normality of generalised least squares estimates, *Econometrica* 52, 811-825.
- Seber, G.A.F., 1977, *Linear regression analysis* (Wiley, New York).
- White, H., 1982, Instrumental variables regression with independent observations, *Econometrica* 50, 483-499.

Table 1
Estimated Mean Squared Errors

Experiment	Estimator	Iterations				
		1	2	3	4	5
1	β_{ls}	0.8170	0.8433	0.8666	0.8859	0.9020
	β_m	0.8266	0.8512	0.8743	0.8936	0.9087
2	β_{ls}	1.5151	1.6075	1.6799	1.7342	1.7738
	β_m	1.4588	1.5883	1.6731	1.7336	1.7765
3	β_{ls}	20.4400	21.9701	23.6265	24.9826	26.0440
	β_m	19.8888	21.8939	23.7231	25.1056	26.1486
4	β_{ls}	31.9520	31.7733	32.7435	33.7856	34.6158
	β_m	27.6794	30.0238	31.9873	33.4609	34.5188
5	β_{ls}	3522.9696	1087.7951	4694.9841	222.9984	4663.7089
	β_m	4014.5814	603.4256	3788.3785	428.9702	4107.2634

UNIVERSITY OF ILLINOIS-URBANA



3 0112 060295927